

APPLICATION FOR UNITED STATES LETTERS PATENT

by

DAVID HOLTZMAN

ROBERT KODEY

and

DAVID POOL

for

**SYSTEM AND METHOD FOR ESTABLISHING AND EVALUATING CROSS
COMMUNITY IDENTITIES IN ELECTRONIC FORUMS**

SHAW PITTMAN
1650 Tysons Boulevard
McLean, VA 22102-5000
(703) 760-7900

Attorney Docket No.: OPI-101-CIP-II

09879220-061301
T06730-0326260

SYSTEM AND METHOD FOR ESTABLISHING AND EVALUATING CROSS COMMUNITY IDENTITIES IN ELECTRONIC FORUMS

RELATED APPLICATIONS

[0001] This application is a continuation-in-part of U.S. Patent Application No. 09/795,968, filed on March 1, 2001, which, in turn, is a continuation-in-part of U.S. Patent Application No. 09/686,516, filed on October 11, 2000, which are both herein incorporated by reference in their entirety.

BACKGROUND

Field of the Invention

[0002] The present invention relates generally to electronic communities where individuals interact and exchange communications over local and world-wide networks. More particularly, the present invention relates to electronic identities and reputations established within such electronic communities.

Background of the Invention

[0003] Electronic communities have been used in the art to facilitate communications between two or more people. Electronic communities typically allow for exchange of information, ideas and opinions over an extended period of time, i.e., a discussion about a particular topic may be initiated by an individual posting a message on day one, and subsequent discussion participants may receive, view or respond to the message at a later date. Electronic communities are similar to non-electronic communities in that members of each electronic community can establish a reputation based on their participation within the community. An electronic community generally provides one or more discussion forums and individual forums may be

dedicated to particular topics. An electronic discussion forum may allow even participants new to the forum to review past discussion messages and therefore to fully participate in the forum. Well-known examples of such communities and electronic forums include Web-based and proprietary message boards (both public and private), USENET news groups, and electronic mailing lists. These electronic communities and discussion forums support both synchronous and asynchronous discussions, i.e., one or more participants may inject communications into the discussion at the same time, or nearly the same time, without disrupting the flow of communications. This allows each individual electronic discussion forum to be rich with communications spanning a wide variety of topics and subjects.

[0004] Other communities and electronic discussion forums may facilitate more traditional asynchronous-like communications by providing, e.g., interactive chat sessions. In these electronic communities and discussion forums, participants are typically online at the same time and are actively responding to messages posted by others. These discussion forums are similar to a traditional telephone discussion in that the information is exchanged in real-time. However, a significant difference is that the electronic discussion forums are, by their nature, written or recorded message transmissions, which may be saved for historical records or for analysis at a future date.

[0005] With so many different topics and subjects within each topic, and so many participants, a significant problem arises in attempting to capture and quantify the communications. Moreover, identifying trends and predicting future behavior in certain markets based on the communications has not been possible in the past

because of the magnitude of the communications and the magnitude of topics and subjects. Further complicating any analysis of communications in electronic discussion forums is the fact that an individual may easily participate in multiple forums by posting the same message in several different discussion forums, and that individuals may use more than one identity when posting.

[0006] Although most electronic communities require each user to select an identity that is unique within a particular community, there has been no coordination among the various communities to allow users to establish a single identity for use within every community. For example, an individual user in the Yahoo.com message boards ("Yahoo community") may have acquired the identity john@yahoo.com. However, because "john" is not very unique, the individual may not be able to use that pseudonym on other communities, such as, e.g., the Amazon.com community. In this example, if the identity john@amazon.com has already been selected by a different individual, then the individual user known as john@yahoo.com would have to select a different pseudonym for use on the Amazon message boards, for example, john2@amazon.com. Essentially, an electronic pseudonym becomes the individual's identity as the user participates in various electronic communities. Thus, this becomes the only way an individual can be referred to within each community or electronic discussion forum.

[0007] The resulting problem for users is a lack of continuity of identity across the various electronic forums they participate in. That is, a single individual cannot easily establish an identity and reputation across electronic communities, even when the forums are related to the same topic. In some instances, a user may prefer such

separation of identities across different electronic communities. For example, a user may wish to participate in one set of communities devoted to financial markets, and another set of communities devoted to building model aircraft. Because there is little relationship between these sets of communities, the user may not desire establishment of a cross-community identity and reputation across both community sets. However, within each set of communities, the user may desire such a cross-community identity. That is, for example, within the various model aircraft communities, the user may wish to build a reputation as a user that provides useful information. Without a way to create a cross-community identity, the user would only be able to establish a plurality of independent reputations, that is, one for each community, with no relationship to each other.

SUMMARY OF THE INVENTION

[0008] The system and method of the present invention allows collection and analysis of electronic discussion messages to quantify and identify trends in various markets. Message information data is collected and becomes a time series stored in a database, indicating the identity or pseudonym of the person posting the message, the contents of the message and other data associated with the message. This data is analyzed to identify when new participants enter and leave the discussion and how often they participate. Calculation of summary statistics describing each community's behavior over time can also be made. Finally, identification of patterns in this data allows identification of pseudonyms who play various roles in each community, as described below.

[0009]

The system of the present invention comprises an electronic discussion data system, a central data store and a data analysis system. The electronic discussion data system may comprise a message collection subsystem as well as message categorization and opinion rating subsystems. The message collection subsystem interfaces with a plurality of pre-determined electronic discussion forums to gather message information. The message categorization subsystem analyzes the message information and categorizes each message according to a plurality of pre-determined rules. Additionally, the message categorization subsystem can perform detailed analysis of the behaviors exhibited by the posting pseudonyms within a community, forum or thread. The opinion rating subsystem further analyzes the message information and assesses an opinion rating according to a plurality of pre-determined linguistic and associative rules. The central data store of the present invention comprises one or more non-volatile memory devices for storing electronic data including, for example, message information, results of analyses performed by the system and a plurality of other information used in the present invention. In a preferred embodiment, the central data store further comprises a relational database system for storing the information in the non-volatile memory devices. The data analysis system of the present invention may comprise an objective data collection subsystem, an analysis subsystem, and a report generation subsystem. The objective data collection subsystem interfaces with a plurality of pre-determined objective data sources to collect data which may be used to establish trends and correlation between real-world events and the communication expressed in the various electronic discussion forums. The analysis subsystem performs the analysis of the objective data

and message information described above. The report generation subsystem generates reports of the analysis to end-users. The reports may comprise pre-determined query results presented in pre-defined report formats or, alternatively may comprise ad hoc reports based on queries input by an end-user of the system.

[0010] The method of the present invention comprises one or more of the steps of collecting a plurality of message information from a plurality of pre-determined electronic discussion forums; storing the plurality of message information in a central data store; categorizing the message information according to a plurality of pre-determined rules; categorizing the behavior exhibited by the pseudonyms within each community, forum or thread; assigning an opinion rating to the plurality of message information based on a plurality of pre-determined linguistic patterns and associative rules; collecting a plurality of objective data from a plurality of objective data sources; analyzing the message information and the objective data to identify trends in the pattern of behavior in pre-determined markets and the roles of participants in electronic discussion forums; and generating reports for end-users of the method based on the results of the analyses performed by the present invention.

[0011] The present invention also provides a system and method for establishing and evaluating cross community identities in electronic communities and discussion forums. The system and method comprise a scheme allowing users to select and register a universal pseudonym which can then be associated with the various local pseudonyms required on each electronic forum. The electronic message postings by a user can be evaluated across the electronic forums to establish a reputation within the communities to which the universal pseudonym relates. This allows the user to

establish a reputation not just within a single electronic forum, but also across multiple communities or forums.

DESCRIPTION OF THE DRAWINGS

[0012] Figure 1 is a schematic diagram of the system architecture employed in a preferred embodiment of the present invention.

[0013] Figure 2 is a schematic diagram of a message collection subsystem implemented in a preferred embodiment of the present invention.

[0014] Figure 3 is a schematic diagram of the hierarchy used to categorize messages in a preferred embodiment of the present invention.

[0015] Figure 4 is an example of graphical report output by a report generation subsystem of the present invention.

[0016] Figure 5 is a schematic diagram of an embodiment of the present invention comprising a pseudonym registration and tracking service.

[0017] Figure 6 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0018] Figure 7 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0019] Figure 8 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0020] Figure 9 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0021] Figure 10 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0022] Figure 11 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0023] Figure 12 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0024] Figure 13 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

[0025] Figure 14 is an exemplary graphical user interface (GUI) page according to an embodiment of the invention.

DEFINITIONS

[0026] **Community** – a vehicle supporting one or more electronic discussions, such as a message board, mailing list, or Usenet newsgroup.

[0027] **Discussion Forum** – an area of a community where discussions directed to a particular theme occur. Examples of discussion forums include the Amazon message board in the Yahoo.com community and the Usenet newsgroup rec.arts.movies.current-films.

[0028] **Message** – the text and associated information posted to discussion forums, also referred to herein as “electronic message”.

[0029] **Topics** – the themes intended for discussion in a discussion forum by a particular community.

[0030] **Subject** – the contents of the “Subject” field in an electronic message posted in an electronic discussion (as distinct from topics).

- [0031] **Discussion Thread** – A series of messages posted within a single forum generally in response to earlier posted messages. Discussion threads typically have the same subject, or were generally created as a “reply to” an earlier message.
- [0032] **Local Pseudonym** – an e-mail address, alias, or other unique handle, i.e., name, used by a participant in an electronic community or discussion forum. A local pseudonym is an end-user’s identity in a particular community.
- [0033] **Universal Pseudonym** – an e-mail address, alias, or other handle used by a participant to associate various local pseudonyms together, enabling the user to establish a cross-community identity. A universal pseudonym is essentially a virtual identity composed of one or more local pseudonyms.
- [0034] **Source** – the issuer of a pseudonym, such as an e-mail host, or the community service provider.
- [0035] **Message Body** – the portion of an electronic message comprising the pseudonym’s contribution to the electronic discussion forum. The message body generally comprises the data, opinions or other information conveyed in the electronic message, including attached documents, files and/or parts of, or entire, previous messages.
- [0036] **Header Information** – the portion of an electronic message not including the message body. Header information generally comprises information related to: the transmission path, time/date stamp, the message poster’s identity, the message identification number (“message ID”), the message subject.

- [0037] **Buzz Level** – for a community or discussion forum, a measure of general activity within the community or forum, as determined by the number of distinct pseudonyms posting one or more messages over a given time frame.
- [0038] **Connectivity** – for a community, a measure of its relatedness with other communities, as determined by the number of other communities in which a community's participants concurrently participate.
- [0039] **Actor** – descriptive name of the role that a pseudonym (local or universal) plays in the social networks of communities. Actors can be further classified according to the following definitions:
- [0040] **Initiator** – a pseudonym that commences a discussion, i.e., one that posts the first message leading to subsequent responses forming a dialog on a particular subject.
- [0041] **Moderator** – a pseudonym that ends a discussion, i.e., one that posts the final message closing the dialog on a particular subject.
- [0042] **Buzz Accelerator** – a pseudonym whose postings tend to precede a rising buzz level in a community.
- [0043] **Buzz Decelerator** – a pseudonym whose postings tend to precede a falling buzz level in a community.
- [0044] **Provoker** – a pseudonym that tends to start longer discussion threads; different from buzz accelerators in that the metric is one discussion thread, not the community's overall discussion level.
- [0045] **Buy Signaler** – a pseudonym whose postings on a topic tend to precede a rising market for that topic.

- [0046] **Sell Signaler** – a pseudonym whose postings on a topic tend to precede a falling market for that topic.
- [0047] **Manipulator** – a pseudonym with little posting history except as manipulators, whose combined postings on one topic elevate the buzz level in the absence of external confirming events.
- [0048] **Connector** – a pseudonym who posts messages related to a large number of different topics or in a large number of different communities.
- [0049] **Mood** – a positive/negative measure derived from analysis of the patterns of actors' behavior.
- [0050] **Topic** - the subject that is being discussed in an electronic community or forum. Many communities have designated one topic per discussion forum. Other communities may designate multiple topics for their forums.
- [0051] **Relevance Score** – a measure of the degree to which a message is relevant to the electronic discussion forum's designated topic or topics. A relevance score may also be assigned to measure the degree to which a message is relevant to a particular thread. An Actor's relevance score measures the degree to which, on average, the Actor posts relevant messages with respect to the discussion forum's designated topic or topics. A discussion forum's relevance score measures the degree to which, on average, the forum's participants post relevant messages with respect to the forum's designated topic or topics.
- [0052] **Impact Score** – a measure of the degree to which a message alters the behavior of other participants within an electronic community, discussion forum, or thread. An Actor's impact score measures the degree to which, on average, the Actor

posts messages with high impact. A discussion forum's impact score measures the degree to which, on average, the forum's participants post messages with high impact.

[0053] **Influence Score** – for individuals, a measure of a pseudonym's potential to affect the views and opinions of other participants within an electronic discussion forum. Similarly, for communities, an influence score is a measure of the degree that recent messages within the community exhibit influence. An influence score is based on a pseudonym's tendency to discuss relevant topics, as well as its impact on that community or discussion. Influence builds on the concepts of relevance and impact by adding a time component, requiring pseudonyms to maintain their influence score over time.

[0054] **Buzz** – a measure of the level of activity within an electronic discussion forum, community, or other group, based on the number of messages that has been posted, weighted by the influence of the authors and relevance of the messages.

[0055] **Migration** – a measure of the movement of pseudonyms and/or influence-weighted posting activity into ("inflow"), or out from ("outflow"), an electronic discussion forum, community, or other group.

[0056] **Flame** – a message determined to be off-topic and emotional.

DETAILED DESCRIPTION OF THE INVENTION

[0057] In a preferred embodiment, the present invention is implemented using a system architecture as shown in Figure 1. The system architecture comprises electronic discussion data system 10, central data store 20, and analysis system 30. Electronic discussion data system 10 interfaces via network 4 with selected electronic

discussion forums 6 to collect electronic messages and analyze intrinsic data comprising the messages according to one aspect of the present invention. Network 4 may be any communications network, e.g., the Internet or a private intranet, and may use any suitable protocol for the exchange of electronic data, e.g., TCP/IP, NNTP, HTTP, etc. Central data store 20 is a repository for electronic messages collected, objective data gathered from external sources and the results of the various analyses or reports produced by the system and method of the present invention. Central data store 20 may be implemented using any suitable relational database application program, such as, e.g., Oracle, Sybase and the like. Data analysis system 30 receives input from selected objective data sources for use in analyzing and quantifying the importance of the electronic discussion messages collected, and provides computer programming routines allowing end-users 9 to generate a variety of predefined and ad hoc reports and graphical analyses related to the electronic discussion messages. Each of the main systems comprising the system architecture of the present invention is described in more detail below.

CENTRAL DATA STORE

[0058] Central data store 20 comprises one or more database files stored on one or more computer systems. In a preferred embodiment, central data store 20 comprises message information database 22, topics database 23, objective data database 24, forum configuration database 25, analysis database 26 and reports database 27, as shown in Figure 1. Message information database 22 comprises the message information collected by message collection subsystem 12. In a preferred embodiment, message information database 22 comprises: a message ID, i.e., a

number or other string that uniquely identifies each message within a community or forum; sender information, i.e., the local pseudonym, e-mail address or name of each message's author; a posting time and date for each message (localized to a common time zone); a collection time and date for each message; a subject field, i.e., the name of the thread or subject of each message; the message body for each message; an in-reply-to field, i.e., the message ID of the message to which each message was a reply; and the source of the message.

[0059] The function and content of central data store 20's database files 23-27 are described in subsequent sections below.

ELECTRONIC DISCUSSION DATA SYSTEM

[0060] As discussed above, electronic discussion data system 10 gathers certain messages and analyzes them according to the intrinsic information comprising the messages. Electronic discussion data system 10 comprises three subsystems: message collection subsystem 12, message categorization subsystem 14 and opinion rating subsystem 16. Message collection subsystem 12 collects message information from data sources and stores the information in central data store 20 for later analysis. Message categorization subsystem 14 extracts information about each message in central data store 20 and categorizes the messages according to a plurality of pre-defined topics. The subsystem analyzes all aspects of each message and determines if the message is relevant to one or more of the topics that the system is currently tracking. A relevancy ranking for each message is stored in central data store 20 for each topic indicating the strength of the message's relation to each topic. Further analysis of the collected message information is carried out by opinion rating

subsystem 16 to determine whether the message conveys a positive, neutral or negative opinion regarding the related topic. Each of the subsystems of electronic discussion data system 10 are described in more detail below.

Message Collection Subsystem

[0061] Message collection subsystem 12 collects electronic message information from the designated electronic discussion forums and passes the collected messages to central data store 20 and to message categorization subsystem 14, as shown in Figure 1. The collected messages comprise records stored in message information database 22 in central data store 20. Database 22 comprises records including message header information and the message body. In a preferred embodiment, each field comprising message header information comprises a separate field of a record in database 22. The architecture used in a preferred embodiment of the present invention for implementing message collection subsystem 12 is shown in the schematic diagram in Figure 2. This architecture supports multiple configurations for data collection and is highly scalable for gathering large or small amounts of message information. Figure 2 illustrates some of the configurations that may be used in a preferred embodiment of message collection subsystem 12.

[0062] As shown in Figure 2, the message collection subsystem consists of several components that function together to collect information from electronic discussion forums 61 and 62 or discussion data files 63 and 64 on distributed networks 41-44. Although shown as separate discussion forums, data files and networks, it would be apparent to one skilled in the art that discussion forums 61 and 63 and data files 63 and 64 could be the same discussion forum or data file, and networks 41-44 could comprise a single distributed network, such as the Internet. Components of message

collection subsystem 12 include message collector programs and message processor programs running on one or more computer systems. The computer systems used by message collection subsystem 12 comprise any suitable computers having sufficient processing capabilities, volatile and non-volatile memory, and support for multiple communications protocols. In a preferred embodiment, the computer systems used by message collection subsystem 12 comprise UNIX-based servers such as available from Sun Microsystems, or Hewlett-Packard and the like. All of the subsystem components can be replicated within a single computer system or across multiple computer systems for overall system scalability.

[0063] In a preferred embodiment, message processor programs, e.g., message processor 121a and 121b, are in communication with database 22, which is part of central data store 20 (not shown in Figure 2). In Figure 2, the message processors and central data store are protected from unauthorized access by firewall security system 122. Other components of message collection subsystem 10 are located at various points in the architecture, as described below. As would be apparent to one of ordinary skill in the art, firewall 122 is provided for security and is not technologically required for operation of the present invention. Message processors 121a and 121b receive information from the message collectors and store the information in the database 22 for later processing. As shown in Figure 2, message processors 121a and 121b may service more than one message collector program to facilitate processing of a large volume of incoming messages. Inbound messages are held in a queue on the message processors, allowing message processors 121a and 121b to receive many more messages from the message collectors than they can

actually process for storing in database 22. This architecture allows the rapid collection of millions of messages from tens of thousands of discussion forums without excessive overloading of the computer systems.

[0064] In a preferred embodiment, the message information collected by message collection subsystem 12 may comprise one or more of the following attributes which are recorded at collection time for use in subsequent analysis by other subsystems of the present invention:

[0065] **Posting date and time** - the date and time the message was posted to the electronic forum. In a preferred embodiment, the date and time indicated by the electronic forum is normalized to reflect the time in a standardized time zone, for example, EST or GMT.

[0066] **Collection date and time** – the date and time the message was collected.

[0067] **Poster's information** – the local pseudonym that posted the message, including any available information such as, e.g., the poster's email address, handle, and community-specific identifiers.

[0068] **Community** – the community in which the message was posted.

[0069] **Forum** - the forum in which the message was posted.

[0070] **Subject** - the subject line of the message, as defined above.

[0071] **Message ID** – the message's unique ID within the community or forum.

[0072] **Body** - the message body as defined above.

[0073] **Message length** – the length of the message, measured in, for example, bytes, characters, lines, or other objective means to indicate message length.

[0074] **Thread** – if a message belongs to a thread, the thread is recorded. In a preferred embodiment, each message's immediate parent, and the original thread parent is stored and is sufficient to reconstruct the thread.

[0075] **Influence score of the local pseudonym** - the influence score of the posting local pseudonym at the time of posting, if one has been previously determined.

[0076] **Reputation score of the universal pseudonym** - the reputation score of universal pseudonym associated with the posting local pseudonym at the time of posting, if one has been previously determined.

[0077] As is known in the art, each discussion forum or data file may have a unique message format. For example, an electronic message from one discussion forum may place the date field first, the message ID second, and the other header and body data last. A different discussion forum may choose to display the message ID first, followed by the local pseudonym of the participant, and the message body. Moreover, each type of discussion forum has its own communications protocol. For example, the communications protocol for an interactive discussion forum (e.g., a chat session) is not the same as the communications protocol for USENET news groups. The message format and protocols need not be static, i.e., as discussion forums evolve, different data structures and protocols may be implemented. To accommodate such changes, each message collector receives configuration information from forum configuration database 25 in central data store 20, either directly or via the message processor systems. The configuration information indicates the data source, i.e., the discussion forum or discussion file, from which messages will be collected. The configuration information further comprises programming instructions tailored for

each individual data source to allow the message collector program to communicate with the data source and extract and parse the message information. Accordingly, message collectors can support a wide variety of protocols utilized by discussion forums including, e.g., HTTP, NNTP, IRC, SMTP and direct file access. In a preferred embodiment, the general programming instructions are written the Java programming language with parsing instructions written in Jpython scripting language. By storing the configuration information in a centralized location, i.e., central data store 20, management of the message collectors is simplified. Accordingly, when the data structure for a particular discussion forum changes, the configuration information needs to be modified only once.

[0078] To ensure compatibility with various computer systems, the message collector programs are written utilizing any suitable programming languages, preferably Java and JPython scripting languages. This allows the collector programs to be easily ported across a wide variety of computer operating systems. Moreover, the message collector programs are designed to have a minimal processing footprint so that they can reside on computer systems that are hosting other critical functions.

[0079] As noted above, there are several ways to implement the architecture supporting message collection subsystem 12. In one implementation, message collector programs, shown in Figure 2 as local message collectors 123a and 123b, are part of local area network ("LAN") 124 and are authorized access through firewall 122. Local message collector 123a interfaces through network 41 to collect messages from discussion forum 61 and local message collector 123b has direct access to discussion data file 63. The latter configuration may be implemented, e.g., if the

operator of message collection subsystem 12 also hosts a community for message discussion forums. As shown in Figure 2, a message collector may collect messages from multiple discussion forums. For example, as shown in Figure 2, local message collector 123b also interfaces through network 41 to collect messages from discussion forum 61.

[0080] In an alternative implementation, message collector programs, such as remote message collectors 125a and 125b, are run on external networks. As shown in Figure 2, the remote message collectors are not part of LAN 124 and do not have direct access to the message processor programs running behind firewall 122. For security reasons, proxy servers 126a and 126b are used to interface with message processor 121b through firewall 122. Functionally, remote message collectors operate in the same manner as the local message collectors. That is, remote message collectors 125a and 125b receive configuration information from central data store 20 (via proxy servers 126a and 126b, respectively). Moreover, remote message collectors may collect messages from discussion forums over a network or directly from discussion data files, as shown in Figure 2. Use of remote message collectors allows for geographic distribution and redundancy in the overall message collection subsystem architecture.

Message Categorization Subsystem

[0081] As known in the art, the actual message topic may not be reflective of the topic assigned to the electronic forum in which a message was posted. Message categorization subsystem 14 analyzes the data collected from discussion forums and categorizes the messages into meaningful groupings, i.e., parent topics and topics, according to predefined rules as described below. In a preferred embodiment,

message categorization subsystem 14 retrieves message information from database 22 and topic information from central data store 20 and stores results of the categorization process in database 22. Alternatively, message categorization subsystem 14 may receive input directly from message collection subsystem 12 for immediate processing into categories.

[0082] Topics database 23 comprises representations of real world topics that are being tracked and analyzed by the system and method of the present invention. Figure 3 shows the hierarchical data structure used in a preferred embodiment of database 23. In a preferred embodiment, abstract root 231, shown in Figure 3 as the top-level of the hierarchy, is not an actual topic stored in database 23 and is shown only to illustrate the hierarchy. Similarly, branches 232-234 are shown in Figure 3 to conceptually show the relationship between topics stored in database 23. Accordingly, branch 232 indicates that some topics stored in database 23 may relate consumer entertainment, branch 233 indicates other topics relate to stock markets, and branch 234 may include other topics, such as, e.g., food, sports, technology adoption, and the like. As shown in Figure 3, the hierarchy comprises one or more parent topics, such as parent topics 235 (related to books), parent topic 236 (related to movies), parent topic 237 (related to market indexes) and parent topic 238 (related to companies). Topics in the hierarchy are the last level, such as, topic 235a (Tears of the Moon), topic 235b (The Indwelling), topic 235c (Hot Six) and topic 235d (The Empty Chair). As shown in Figure 2, topics 235a-235d are related to each other by parent topic 235 (books).

[0083] In a preferred embodiment of the present invention, message categorization subsystem 14 assigns a relevance score for each topic to each message collected by message collection subsystem 12. The relevance score is determined based on a set of predefined rules stored in database 23 for each topic. The rules comprise a series of conditions defining information relevant to the topic, having an associated weighting to indicate the strength a particular condition should have in determining the overall relevance rank of the message with respect to a topic. Messages that need categorization are processed by message categorization subsystem 14 synchronously, i.e., the rules for each topic are applied to each message regardless of the relevance score for prior topics analyzed. The elements of each message, including e.g., subject, source, and content are processed against the conditions of each topic in the database. Based on the conditions that are satisfied and the weights of those conditions, a relevance score for each topic is assigned to each message. As messages are processed, their relevance score for each topic is updated in message information database 22 in central data store 20. Relevance scores are described herein after in greater detail.

[0084] An example of the rules which may be processed by message categorization subsystem 14 is presented in Table 1, below. In this example, the topic is “The Perfect Storm” which, as shown in Figure 3, is under the parent topic “Movies.” The conditions for determining the relevance ranking for each message in this example are shown in Table 1, below.

| <i>Condition</i> | <i>Weight</i> |
|--|---------------|
| Message originated from Yahoo movie discussion forum. | 10 |
| Message subject contains "The Perfect Storm" | 90 |
| Message subject contains "Perfect Storm" | 80 |
| Message body contains "The Perfect Storm" | 50 |
| Message body contains "The Perfect Storm" and "George Clooney" | 90 |
| Message body contains "Warner Brothers" and "Barry Levinson" | 75 |

Table 1

[0085] The number, nature and weights for conditions used to determine the relevancy ranking for each topic depends on the nature of the topic itself. The accuracy of the relevancy ranking assigned can be increased by refining the conditions and weights after analysis of the results obtained by the system. For example, analysis of the results in the above example may show that an additional condition, such as "Message originated from Yahoo movie discussion forum and message subject contains "Perfect Storm"" should be included in the rules and have a weight of 99. If subsequent analysis provides refined rules, message categorization subsystem 14 may be re-run against each message in database 22 to update the relevancy rankings, if desired.

[0086] In addition to determining the actual message topic for a message, message categorization subsystem 14 may compute additional message attributes such as:

[0087] **Thread length** – the number of messages in the thread the message belongs to at the time the attribute is computed. This attribute can change over time and if computed, should be periodically updated to reflect new messages posted to the forum.

[0088] **Position in thread** – the message’s position within its thread. Position could be expressed as a location, e.g., first, second, third, etc., message in the thread or some other expression reflecting the order of message’s occurrence in the thread.

[0089] **Relevance score** – an indication of whether the message is truly relevant to the intended topic, i.e., whether the message’s actual topic is related to the forum’s designated topic or in the case of a thread, the thread’s topic. The actual message topic and the strength of the score used to determine the topic, as described above, are used to establish a relevance score. In a preferred embodiment, the relevance score is computed as a numeric value from 0.0 to 1.0, with a score of 0.0 indicating no connection between the message’s actual topic and the forum’s (or thread’s) topic and a score of 1.0 indicating the message is fully relevant. Because a particular forum may have multiple topics, more than one relevance score may be computed. In a preferred embodiment, the message is assigned the highest computed relevancy score.

[0090] In a preferred embodiment, the relevance score for a message is computed using a textual pattern analysis software, such as available from Verity Inc, of Mountain View, CA.

[0091] **Impact score** – an indication of the message’s impact on the discussion forum. In a preferred embodiment, the impact score is computed as a numeric value from 0.0 to 1.0, with a score of 0.0 indicating the message had no impact on discussion behavior and a score of 1.0 indicating the message had great impact on the discussion. In one embodiment, the impact score can be based on the rate of new postings to the forum immediately following the posting of the message compared to the rate of new posting immediately prior the posting of the message. In another

embodiment, the impact score measures changes in the number of pseudonyms participating in a discussion after the message has been posted versus before. In a preferred embodiment, irrelevant, or off-topic messages i.e., messages with a low relevance score, receive an impact score of zero, while highly relevant, or on-topic messages i.e., messages with a high relevance score, receive two impact scores: “impact score” I , measures impact using all messages posted in the forum; and “relevant impact score” I_r , measures impact using only relevant messages posted in the forum. In another preferred embodiment, the impact score is measured not only on the change in message traffic experienced in the forum, but also may incorporate changes in the number of threads, the reputation assigned to replying pseudonyms, changes in message vocabulary and style, and changes in topics for messages posted to the forum after the message has been posted.

[0092] In a preferred embodiment, a message’s impact scores can be computed as follows. For every message, compute window of time T that the message’s impact will be measured over. For a given message m , T is the amount of time it took for p unique pseudonyms to post a message before the current message, excluding the poster of m . Next, determine P_a , which is the number of unique pseudonyms that post a message during time T after m . Next, determine P_b , which is the number of unique pseudonyms that post a message during time T before m . Using these values, the impact score, I , for the given message is:

$$I = \frac{|P_a - P_b|}{|P_a + P_b|}$$

[0093] Similarly, in a preferred embodiment, a message m ’s relevant impact score, I_r , is:

$$I_r = \frac{|P_r a - P_r b|}{|P_r a + P_r b|},$$

where, $P_r a$ is the number of unique pseudonyms that post a relevant message during time T_r after m , $P_r b$ is the number of unique pseudonyms that post a relevant message during time T_r before m , where T_r is the amount of time it took for p unique pseudonyms to post a relevant message before the current message, excluding the poster of m .

[0094] In a preferred embodiment, times T and T_r are bounded by a predefined minimum and maximum to keep the calculations stable. Without such bounding, extremely active forums could accumulate p unique pseudonyms so fast that the results we could be very volatile. On the other hand, extremely inactive boards could take a very long time to accumulate p unique pseudonyms.

[0095] Example: Table 2 below illustrates a representative set of messages posted to an electronic discussion forum. In the table, the current message, c , was posted at time = 11, and was posted by a user with the local pseudonym "A4." Applying the above formulas for $p = 3$, I and I_r can be calculated for message m as follows:

[0096] a) $T = 3$ time units (the first unique pseudonym, A2, posted a message at time = 10, the second unique pseudonym, A4, posted a message at time = 9, and the third unique pseudonym, A3, posted a message at time = 8, so it took from time = 8 to time = 11 to get three unique pseudonyms prior to message c), $Pa = 3$ unique pseudonyms (during the three time units following message m , three unique pseudonyms, A3, A4 and A2, posted messages), $Pb = 2$ unique pseudonyms (during the three time units before message m , only two unique pseudonyms, A3 and A2, posted messages), therefore $I = 1/5$.

community exhibit influence. In a preferred embodiment, influence and relevant influence scores are set to decay over time so that pseudonyms that stop posting messages in electronic forums will loose their influence over time.

[0099] In a preferred embodiment, each pseudonym receives an influence score, F , based on the impact and relevance scores for the messages that they author. In a preferred embodiment, influence is computed daily, and is based on historic message scores for the set of messages authored by a given pseudonym. Also, in a preferred embodiment, influence scores will be decayed according to the following function:

$$d = e^{\left(\frac{t_m - t}{\tau}\right)}$$

where t_m is the date and time the message was posted, t is the current system date and time, and τ is a configurable constant that controls the rate of decay. For a given value of τ , there is a maximum $(t_m - t)$, that will be considered significant, such that the result of the decay function above is ≥ 0.001 .

[0100] For a given pseudonym, the influence score, F , is:

$$F = a \cdot \sum_{i=1}^n (Rel_i \cdot d_i) + b \cdot \sum_{i=1}^n (I_i \cdot d_i)$$

where n is the number of messages authored by the pseudonym that were posted within the influence window, Rel is the relevance score for a message, I is the impact score of a message, d is the time decay function for a message, as defined above, and a and b are configurable constants that control the weightings of relevance and impact.

[0101] Similarly, for a given pseudonym, the relevant influence score, F_r , is:

$$F_r = a \cdot \sum_{i=1}^n (Rel_i \cdot d_i) + b \cdot \sum_{i=1}^n (I_{ri} \cdot d_i)$$

[0102] where n is the number of messages authored by the pseudonym that were posted within the influence window, Rel is the relevance score for a message, I_r is the impact score of a message, d is the time decay function for a message, as defined above, and a and b are configurable constants that control the weightings of relevance and impact.

[0103] Moreover, in a preferred embodiment, a community or forum may be assigned an influence score F , which is computed in the same way as for pseudonyms, where the measured message set is defined by the messages that belong to the given community or forum.

[0104] **Reputation score** – a measure of the reputation a particular universal pseudonym possesses within a community, forum or thread. The reputation score, R , assigned to a pseudonym is a function of the influence scores assigned to local pseudonym associated with the universal pseudonym. The relevant reputation score, R_r , assigned to a pseudonym is a function of the relevant influence scores assigned to local pseudonym associated with the universal pseudonym. In a preferred embodiment, reputation scores are set to decay over time so that if a user with a universal pseudonym stops posting messages with associated local pseudonyms in electronic forums, reputation will be lost over time.

[0105] In a preferred embodiment, a universal pseudonym's reputation score, R , is computed as follows:

$$R = \frac{\sum_{i=1}^P (F_i \cdot n_i)}{\sum_{i=1}^P (n_i)}$$

where P is the number of local pseudonyms associated with the universal pseudonym, F is the influence score for a given local pseudonym, and n is the number of messages used to compute the local pseudonym's influence score.

[0106] Similarly, a universal pseudonym's relevant reputation score, R_r , is computed as follows:

$$R_r = \frac{\sum_{i=1}^P (F_{ri} \cdot n_i)}{\sum_{i=1}^P (n_i)}$$

where P is the number of local pseudonyms associated with the universal pseudonym, F_r is the relevant influence score for a given local pseudonym, and n is the number of messages used to compute the local pseudonym's influence score.

[0107] **Leadership score** – a measure of a particular pseudonym's tendency to lead or follow a discussion within a forum or thread. The leadership score, L , assigned to a pseudonym can be thought of as a measure of the degree to which a pseudonym is an early participant in current discussions in the forum. Similarly, the relevant leadership score, L_r , assigned to a pseudonym can be thought of as a measure of the degree to which a pseudonym is participating in current discussions in the forum by posting relevant messages. In a preferred embodiment, the number of threads that a pseudonym posts to is weighted more heavily than the pseudonym's raw number of posts in the discussion forum. This removes the effect of one-on-one and repetitive conversations within threads which may generate substantial message traffic, but does

not lead to greater involvement among other group participants. The leadership score can be assigned based on a variety of factors, such as, the number of threads in which the pseudonym participates, the location of the pseudonym's postings in the threaded discussion, i.e., the earlier the pseudonym posts messages in a threaded discussion, the higher the leadership score will be.

[0108] In a preferred embodiment, a pseudonym's leadership score, L , is the sum of the minimum location in each thread posted to, divided by the sum of the length of each thread posted to:

$$L = \sum_{i=1}^P \frac{\text{Min}_i}{T_i}$$

where for each thread in the forum for which the pseudonym posts messages, Min is the location in the thread of the pseudonym's earliest posting, T is the length of the thread, and P is the number of threads being considered in the computation.

[0109] Similarly, in a preferred embodiment, a pseudonym's relevant leadership score, L_r , is computed as follows:

$$L_r = \sum_{i=1}^P \frac{\text{Min}_{ri}}{T_{ri}}$$

where for each thread in the forum for which the pseudonym posts messages, Min_r is the location in the thread of the pseudonym's earliest relevant posting and T_r is the length of the thread.

[0110] Finally, message categorization subsystem 14 may compute aggregated attributes for groups (also referred to herein as "sets") of messages within a community, topic, forum or thread. A set could also comprise a group of messages posted in a forum by a single local pseudonym, or a group of messages posted on

multiple forums by a user having a universal pseudonym with associated local pseudonyms on each forum. Such aggregate attributes include for example the following:

- [0111] **Last post date and time** – date and time of the last post in the set.
- [0112] **Number of posters** – total number of pseudonyms used by those posting messages in the set.
- [0113] **Distribution of posts** – a breakdown of the distribution of posters and messages in the set, i.e., an indication of whether or not most of the messages in the set come from a small number of pseudonyms.
- [0114] **Number of posts** – the total number of posts in the set.
- [0115] **Number of threads** – the number of unique threads (or subjects, if threading is unavailable) within the set of posts.
- [0116] **Set relevance score** – an indication of number of messages found to be relevant, i.e., on topic, for the forum in which they were posted. In a preferred embodiment, the set relevance score is determined by computing the average relevance score of messages in the set, for a given date or number of messages.
- [0117] **Set impact score** - aggregate impact and relevant impact scores can be determined by computing the average impact or average relative impact of the specified message set. For more granularity, these scores can be computed for messages posted on a given date or for a number of messages selected from the set.
- [0118] **Set flame score** – the percentage of messages found to be not relevant, i.e., not on topic, for the forum in which they were posted.

periodically on day n , such as hourly. Generally, if the computed M is positive, it is reported as an “inflow.” If the computed M is negative, it is reported as an “outflow.” As with other attributes, the migration score can be computed as a relevant migration score based on the relevant raw buzz score, B_r .

Tracking of Buzz Scores and Migration Scores in an Exemplary Embodiment

[0129] Raw buzz scores, relative buzz scores, and migration scores can be aggregated over a set of messages that are derived or selected in various fashions. For example, the set of messages can correspond to a community, multiple communities, a forum, multiple forums, a topic, multiple topics, a thread, multiple threads, a pseudonym, or multiple pseudonyms (including a universal pseudonym).

[0130] According to one embodiment, electronic discussion system 10 (Figure 1) tracks a series of topics for purposes of financial analysis. For example, electronic discussion system (EDS) 10 might track the following indices: NASDAQ 100, Dow Industrials 30, and the S&P 500. EDS 10 may also track the ten sectors of the Dow Jones Industry Classification System. EDS 10 might also track the individual stocks that comprise the aforementioned indices and sectors. This embodiment is hereafter referred to as the “financial embodiment.” Of course, this series of topics is exemplary only, and EDS 10 could easily be configured to track other series of topics, such as all stocks listed with the SEC, all movies currently being shown at theatres, various television programs, and so forth.

[0131] EDS 10 may periodically (e.g., each hour) perform collection jobs from a series of discussion forums in order to collect messages that may be relevant to the series of topics. In the financial embodiment, EDS 10 may collect messages from the

following online financial communities: Raging Bull, Motley Fool, Silicon Investor, and Yahoo. Other financial communities could easily be used for the collection jobs. These electronic discussion forums (i.e., entire communities or forums/groups/boards within a community) that are accessed by EDS 10 may be referred to as the “sources.”

[0132] Preferably, central data store 20 (Figure 1) includes a topics database 23 that associates each source with one or more topics. For example, topics database 23 may include a “board-topic link” (BTLINK) table that lists each of the boards (discussion groups) in the source communities from which messages are collected. The topics that are associated with each board are provided in the BTLINK table. Some boards may be associated with a single topic (e.g., the Amazon board on Yahoo is associated with the stock topic Amazon), while other boards may be associated with multiple topics (e.g., the rec.arts.movies.current-films board on Usenet is associated with all movie topics in the system). A BTLINK table can be used so that EDS 10 processes each message for the proper topic(s).

[0133] For each message collected, EDS 10 may compute a relevance score for one or several topics that are associated with the source. EDS 10 may also compute an impact score and/or relevant impact score for the message. If a relevant impact score is computed, one relevant impact score may be computed for each topic associated with the source.

[0134] In a preferred embodiment, only the highest relevance score for each message is used. For example, assume that message X is collected from Board Y. The BTLINK table indicates that Board Y is related to Stock Topic 1 and Stock Topic 6.

The relevance score for message X is higher for Stock Topic 6 than for Stock Topic 1. The relevance score for Stock Topic 6 will be used, to indicate that message X is most relevant to Stock Topic 6, and the other relevance score is discarded. If a relevant impact score is needed, it is only computed for Stock Topic 6.

[0135] Based on the collected message information, EDS 10 can extract the pseudonym author of the message. Preferably, a so-called “author’s table” is maintained in central data store 20 and may be used to track historic author scores, such as relevance, impact, influence, and the like. The author’s table contains identifiers for the pseudonyms that have previously been tracked. Preferably, the author’s table assigns a track number to each such pseudonym.

[0136] The extracted pseudonym author is compared to existing data or otherwise analyzed to determine if the author is a previously tracked author or a new author. Preferably, the extracted pseudonym is compared to the author’s table to make this determination.

[0137] If the extracted author is a new author, EDS 10 may update the author’s table to include the new posting pseudonym. Preferably, a new track number is assigned to the new author.

[0138] If the extracted author is not a new pseudonym, EDS 10 may compute an updated influence score and/or relevant influence score based on the new message posted by the extracted author. EDS 10 may also compute an updated reputation score and/or relevant reputation score based on the new message. Of course, updates to cumulative measures such as influence and reputation need not be performed on a message-by-message basis. Such updates could be performed on a less regular basis,

such as by performing a periodic batch job involving many messages. Data for influence scores and reputation scores may be stored in the author's table or at a separate database.

[0139] Preferably, collection jobs collect messages from the sources on a periodic basis, such as hourly. Multiple messages are typically collected in each collection job, so that computation of relevance scores, impact scores, influence scores, and reputation scores can be performed in an efficient, batch-type fashion.

[0140] Periodically, such as on an hourly basis, EDS 10 computes buzz scores for the series of topics that are tracked. In the financial embodiment, buzz scores for individual stock topics can be computed directly from the relevance scores (from the messages) and the influence scores (for the posters of the messages) associated with that topic. Buzz scores for "level-up topics," such as sectors and indices, are computed by "rolling up" or aggregating the buzz scores for the topics making up the level-up topics. For example, a buzz score for the S&P 500 is computed by aggregating the buzz scores for the stock topics making up the S&P 500. The buzz score for the Dow Technology Sector is computed by aggregating the buzz scores for the stock topics making up the Technology Sector.

[0141] Preferably, EDS 10 computes both raw buzz scores and relative buzz scores. The relative buzz scores can be analyzed to identify unusual discussion patterns. According to one embodiment, the relative buzz scores can be compared to a so-called "anomaly threshold" in order to flag unusual buzz patterns. In the financial embodiment, the relative buzz score may be compared to a threshold (e.g., a threshold of about 2-5, preferably about 3) in order to identify companies that are "in play" or

that are otherwise receiving an unusual relative level of attention by influential posters.

[0142] Periodically, EDS 10 also computes the migration scores for at least some of the topics that are being tracked. For example, in the financial embodiment, EDS 10 may compute the migration scores for each of the DOW sectors. Migration scores for the DOW sectors can be presented together in order to provide insight into the migration across sectors. Individual migration scores can be presented in order to provide insight into migration by sector over time. Providing migration scores based on sector is an exemplary embodiment. Migration scores could be provided for any or all of the series of topics tracked by EDS 10.

[0143] Figures 6-14 provide exemplary graphical user interface (GUI) display pages that could be provided according the financial embodiment. In Figure 6, an introductory page (“Dashboard”) provides an index buzz graph 605 and top-in-play graph 610. The in-play stocks are displayed in in-play area 615. Preferably, the in-play stocks may be selected as those stock topics having relative buzz scores exceeding an anomaly threshold of about 3. Sector migration list 620 displays the sectors with the top inflow migration scores. Link 625 provides links to the online communities that are the sources of EDS 10 (Figure 1). Topic selection area 625 can be used to select other topics for display on the introductory page. Preferably, topic selection area 625 permits the user to select stock topics (“Symbol”) or sector topics (“Sector”).

- [0144] By selecting the sector topic “Technology” in topic selection area 625, the display page of Figure 7 is presented. In Figure 7, the upper graph now displays a relative buzz graph 705 for the Technology sector topic.
- [0145] By selecting the stock topic “QCOM” in topic selection area 625, the display page of Figure 8 is presented. In Figure 8, the upper graph now displays a relative buzz graph for the QCOM stock topic.
- [0146] By selecting software button 630 (“Watch List”) in Figure 6, the display page of Figure 9 is presented. The Watch List is a programmable function for a user to select specific topics that interest the user. In Figure 9, watch list area 905 lists several stock topics and a sector topic that interest the user. The user can delete topics from the list using delete area 910 and the user can add topics using add area 915.
- [0147] By selecting software button/link 635 (Figure 6), the user can access a detailed display page such as that presented by Figure 10. The exemplary display page of Figure 10 is a two-pane display enabling comparisons between topics, such as between a first stock and a second stock, a first stock and a sector, and so forth. Topic input area 1005 allows input of a stock topic or sector topic. Source input area 1010 allows the user to select the online communities to be used (e.g., Yahoo, Raging Bull, All Communities, etc.) for the computation. Market events input area 1020 allows the user to select what external market indicators are to be viewed, such as volume, price, price range, average trade size, relative strength index, on-balance volume, and so forth. Time input area 1025 allows the user to select the timeframe, such as one year, six months, three months, one week, one day, and so forth. Computation input area 1015 allows the user to select the scoring that is to be viewed, such as raw buzz,

relevant raw buzz, relative buzz, relevant relative buzz, and so forth. Initiate button 1030 initiates a computation.

[0148] In the exemplary display for Figure 10, the comparison between the relative buzz and volume for the stock topic “BVSN” is displayed in first area 1035. Second area 1040 displays the relative buzz for the “Technology” sector topic as a whole. Meter display 1045 provides a meter-type display of the relative buzz. Preferably, meter display 1045 displays the current relative buzz, the maximum relative buzz for the day (referred to as the “high water” relative buzz), and the anomaly threshold (3).

[0149] If the user changes “relative buzz” to “buzz” in computation input area 1015 (Figure 10), the exemplary display page of Figure 11 may be presented. In Figure 11, first area 1105 and second area 1110 now display graphs of raw buzz for the stock topic and sector topic, respectively.

[0150] If the user changes “volume” to “price” in market events input area 1020 (Figure 10), and the user also changes the “Technology” sector topic to the “IBM” stock topic in topic input area 1005 (right hand pane), the exemplary display page of Figure 12 may be presented. First area 1205 now displays a graphical comparison between the relative buzz and the price of the BVSN stock topic. Second area 1210 now displays similar information for the IBM stock topic.

[0151] If the user selects software button/link 640 (Figure 6), an exemplary migration scoring display page such as that of Figure 13 may be presented. In Figure 13, a first area 1305 presents migration across sectors. A second area 1310 presents a graph of migration by sector. The timeframe and sector can be selected using time input area

1315 and sector input area 1320, respectively. List 1325 and list 1330 may provide lists of top inflow sectors and top outflow sectors, respectively.

[0152] If the user selects software button 650 (Figure 6), a top relative buzz display page such as that of Figure 14 may be presented. Figure 14 includes a list 1405 of topics having the top relative buzz scores. List 1405 may include a first list for stock topics 1410 and a second list for sector topics 1415.

Opinion Rating Subsystem

[0153] Opinion rating subsystem 16 extracts message information from database 22 in central data store 20 and assigns an opinion rating for each message by analyzing textual patterns in the message that may express an opinion. The textual patterns are based on linguistic analysis of the message information. For example, if the message body includes words such as “movie” and “awful” in the same sentence or phrase and the message had a high relevancy ranking for the topic “The Perfect Storm” the message may be expressing a negative opinion about the movie. Textual pattern analysis software, such as available from Verity Inc, of Mountain View, CA, may be used to assign the opinion rating for each message. Such passive opinion polling is useful for market analysis without the need for individually interviewing active participants in a survey. Once the rating process is complete, the rating for each opinion processed is stored in database 22 in central data store 20. In a preferred embodiment of the present invention, opinion rating subsystem 16 computes sentiment scores for each message as described below.

[0154] **Sentiment score** - measures the degree to which a message exhibits positive or negative sentiment. The definition and measure of positive and negative sentiment may change from product to product, since the way people convey their opinions is

very different depending on the subject being evaluated. For example, in the financial world, positive sentiment may be measured by detecting “buy” signals from the author. Negative sentiment may be measured by detecting “sell” signals. For entertainment, positive sentiment may be measured by detecting phrases that convey a general liking of the subject. Negative sentiment may be measured by detecting phrases that convey a general disliking of the subject.

[0155] As described above, sentiment is based on the message content, not just header information. In a preferred embodiment, positive sentiment and negative sentiment scores are individually computed for every message. Each result is bounded 0.0 to 1.0, where 0 indicates no sentiment was detected, and 1 indicates full positive and/or negative sentiment. A message may receive both positive and negative sentiment scores. Individual sentiment scores for a particular message are also referred to as “raw sentiment” scores. Raw sentiment scores may be either positive raw sentiment scores or negative raw sentiment scores. Aggregate sentiment scores can be computed based on aggregate raw sentiment scores for a community; for a particular subject in a community, or for a pseudonym.

[0156] The elements used to compute raw sentiment in a preferred embodiment include: message subject, message text; and positive and negative sentiment models. Similarly, the elements used to compute aggregate sentiment scores include sentiment scores of messages to be aggregated; and optionally includes date information for the messages and a total counting the number of messages in the aggregation.

[0157] In a preferred embodiment, a Verity engine will be used to score both positive and negative raw sentiment for each message according to the input sentiment model.

The Verity output can be stored in a categorization table in a database such as Oracle, and may consist of a score from 0.0 to 1.0 for every message/sentiment combination in the query set. Aggregation may be achieved by computing the average score of a specified message set, for a given date or number of messages. The message set is defined as the set of messages that belong to the specified subject, community, and/or pseudonym.

[0158] The aggregation, or “average score” above, can take many forms. Positive and negative sentiment may be treated as independent to compute an aggregate “positive sentiment” and an aggregate “negative sentiment” score. Or, an aggregate “net sentiment” score can be computed by combining positive and negative raw sentiment scores. The equations below may be used to computer aggregate sentiment scores for a set of messages:

$$S_p = \sum_{i=1}^m s_{pm}$$

$$S_n = \sum_{i=1}^m s_{nm}$$

$$S = \sum_{i=1}^m (s_{pm} - s_{nm})$$

[0159] Where S_p is the aggregate positive sentiment, S_n is the aggregate negative sentiment, S is the aggregate net sentiment, m is the set of messages, s_{pm} is the positive sentiment score for a particular message, and s_{nm} is the negative sentiment score for that message.

[0160] Aggregation for a subject or community can be performed for all messages posted within a given date. Further. intraday aggregation may be achieved by setting

the target date to the current system date, and will include only those messages that have already been collected and categorized for the day.

[0161] In a preferred embodiment, sentiment (which relates to opinions expressed in a message) and relevance scores (which relates to subject matter addressed in the message) are measured completely independent of one another. The two scores (sentiment and relevance) may be further correlated for the purpose of detecting sentiment about a specific topic. In one embodiment of the present invention, such correlation is computed by using the highest relevance score (what is the message most about) and the highest sentiment score (is the message mostly positive or negative). In this manner a sentiment score may be determined for a particular topic. Additionally, topical aggregate sentiment scores may be computed by aggregates sentiment scores for messages directed to a common topic.

DATA ANALYSIS SYSTEM

[0162] Data analysis system 30 comprises objective data collection subsystem 32, analysis subsystem 34 and report generation subsystem 36, as shown in Figure 1. The overall goal of data analysis system 30 is to identify and predict trends in actual markets based on the electronic discussion data being posted to various electronic discussion forums and to provide reports for end-users 9 of the system and method of the present invention.

Objective Data Collection Subsystem

[0163] Objective data collection subsystem 32 collects objective data from both traditional and electronic sources and stores the information in database 24 on central data store 20 for later analysis. Objective data sources 8, shown in Figure 1, may include for example, market data such box office sales for recently released movies,

stock market activity for a given period, television viewer market share (such as Nielsen ratings), and other such objective data. The specific data collected from each objective data source depends on the nature of the market being analyzed. For example, objective data on the stock market may include: a company's name; its Web home page address, i.e., universal resource locator; ticker symbol; trading date; opening price; high price; low price; closing price and volume. In other markets, the objective data may include: sales, measured in units sold and/or revenue generated; attendance at events; downloads of related software and media files; press release date, time and key words; news event date; and the like. The objective data is used by analysis subsystem 34 to identify and predict trends and correlation between real world events and electronic discussion data, as described below.

Analysis Subsystem

[0164] Analysis subsystem 34 performs analysis of the information collected by the message collection subsystem 12 and objective data collection subsystem 32, and the categorization and opinion information determined by message categorization subsystem 14 and opinion rating subsystem 16, respectively. Analysis subsystem 34 determines the existence of any correlation between discussion forum postings and market activity for each topic that the system is currently tracking. The results of the analysis are stored in the analysis database 26 in central data store 20 for eventual presentation to end-users 9. Analysis subsystem 34 examines the internal behavior of communities and correlates individual and group behavior to the world external to the communities using a variety of analysis techniques with a variety of goals. Analysis subsystem 34 identifies and categorizes actors by measuring the community's response to their postings; measures and categorizes the community's mood;

correlates actors' behavior and the communities' moods with objective data sources; and forecasts the markets' behavior, with confidence estimates in various timeframes. Identifying and tracking both the actors and the community mood is important, because the effect of an actor's message depends in part on the mood of the community. For example, an already-nervous community may turn very negative if a buy signaler or other negative actor posts a message, while the same message from the same person may have little effect on a community in a positive mood. The following sections describe the patterns sought in the analysis and describes how the community behaves after postings by each local pseudonym associated with the patterns.

a. Actor Classification

[0165] Actors are classified by correlating their postings with objective data, which is external to the electronic forum. Changes in the objective data (e.g., stock price changes, increased book sales, etc.) are tracked during several discrete short time periods throughout a longer time period, such as day. A score is assigned to each local pseudonym posting messages related to a given topic based on the change observed in the objective data from the preceding discrete time period. A local pseudonym's score may be high, medium or low, depending on the magnitude of the change. For example, in a preferred embodiment, local pseudonyms who tended to post messages just prior to major increases in stock price, receive a high positive scores; while those whose postings tended to precede major drops have the lowest negative scores. The scores assigned to a local pseudonym during the longer time period are aggregated into a composite score for the local pseudonym.

results output by analysis subsystem 34. Provokers are local pseudonyms that tend to start longer discussion threads, which may contribute to a community's overall discussion level, but is not indicative of a rise in discussion level for the community. Again, identification and tracking of provokers allows better results in the analysis of electronic discussion information. Finally, a connector is a local pseudonym that posts on a high number of topics or a high number of communities.

[0169] Analysis subsystem 34 tracks and observes the behavior characteristic of the local pseudonyms posting messages to electronic discussion forums and assigns a reputation score indicating their categorization. In a preferred embodiment, the reputation score comprises an array of ratings for each of the possible categorizations. From the reputation score, composite views of the tendencies of the local pseudonyms can be formed to graphically illustrate the local pseudonym's reputation in a given community. An example of one such composite view is shown in Figure 4, wherein a local pseudonym's reputation as a buzz accelerator/decelerator is plotted against its reputation as a buy/seller signaler. As shown in Figure 4, local pseudonym A has a strong tendency as a buy signaler and is a buzz accelerator, but not a strong buzz accelerator. In contrast, local pseudonym B has strong tendencies as both a sell signaler and a buzz decelerator in the market. The impact of the classifications depends, of course on the market involved, as discussed previously.

b. Community Mood

[0170] As discussed above, a local pseudonym's classifications are useful to the extent they can quantify the tendencies of the various actors in a community. However, the impact of such actors on the community depends not only on the tendencies of the actors, but also on the overall mood of the community. The measure

of a community's mood is determined from the change in discussion levels in the community. The mood assigned is based on observed trends for the associated topic. For example, when discussion levels rise in stock market forums, the rise is usually accompanied by a drop in stock market prices due to increased selling activity, indicating a negative mood in the community. Similarly, an increase in discussion levels for a movie topic may indicate a generally positive mood for the community. Other indicators of community mood include the number of new participants in a community, which correlates to an increased interest in the community's topic. Moreover, the combined positive and negative influence scores of actors in a community is an indicator of its overall sentiment. Another factor indicating a community's mood is its turnover rate, i.e., the number of new participants versus the number of old participants, indicates the depth of interest in the community's topic.

[0171] The combined provocation-moderation scores of active participants is expected to be a forecaster of the community's discussion near-term discussion level.

[0172] The ratio of message volume to external volume (stock trading volume in the prototype) will be explored as an indicator of confidence for other forecasts.

[0173] The number of active discussion threads, relative to the number of participants, is an indicator whose significance we plan to explore. "Flame wars," for example, are typically carried out by a small number of people generating a large volume of messages.

[0174] The ratio of "on-topic" to "off-topic" messages, which we expect to be able to measure via linguistic analysis, is an indicator whose significance we plan to explore.

[0175] Co-occurrence of topics within a community, also measurable via linguistic analysis, is an indicator of shared interests among communities, whose significance we plan to explore.

c. Algorithms and Modeling

[0176] As discussed above, the analysis system uses patterns in message postings to identify community moods and opinion leaders, i.e., those local pseudonyms whose postings can be correlated to changes in the market and/or forum discussion levels. Linguistic analysis extends this analysis by showing and summarizing the subjects under discussion and reveals attitudes toward the topics discussed. The linguistic analysis used in the present invention is not intended to explicitly identify any individual's attitude toward a given topic; rather the overall attitude of the community is assessed.

[0177] The analysis system relies on the inherent repeated patterns in discussions that yield accurate short-term forecasts. The existence of such repeated patterns is known in the art, and can be explained with reference to three areas of research into social networks. *Chaos and complexity theories* have demonstrated that large numbers of agents, each of whom interacts with a few others, give rise to repeating patterns by virtue of simple mathematics. *Social network theory* grounds mathematical models in human behavior. *Computer-mediated communications* research applies the mathematical models to "new media" technologies including the Internet.

[0178] As with any high-frequency, high-volume data mining challenge, the number of potential variables is enormous and the applicable techniques are many. To simplify this problem, the system and method of the present invention reduces the data sets as much as possible before analysis. Accordingly, on the assumption that

there are a very small number of opinion leaders relative to participants, the vast majority of participants whose postings did not occur near objective data inflection points, i.e., sharp changes in the objective data, are eliminated. This greatly reduces the amount of data that is further analyzed by the system and method of the present invention. The period of time over which inflection points are identified has a great impact on which patterns can be identified and usefulness of the resulting data. For example, stock price movement and other markets are known to have fractal patterns, so they have different inflection points depending on the time frame chosen. Accordingly, different inflection points will be identified if the period is weekly, monthly, or yearly. The more volatile a market is, the more inflection points can be found.

[0179] The following sections describe the various types of analyses used in a preferred embodiment of analysis subsystem 34.

Statistical Analysis

[0180] Histograms divide scores into “bins” that show the distribution across the range of values. Histograms of the positive/negative influence scores, as well as the provoker/moderator scores described above, are used to select statistically significant local pseudonyms at the outlying ends of the normal distribution curve. A database query can then calculate the ratio of these opinion leaders who have posted in the last X days. For example, if 25 of the top 50 “positives” and 10 of the top 50 “negatives” posted in the last two days, the ratio would be 2.5, indicating that positive market movement is more likely than negative.

(1) Fourier Analysis

[0181] Fourier analysis is a well-established technique, with many variations, for breaking down a complex waveform, such as plots of discussion levels, into component waves. This makes it possible to subtract regularly occurring waves, such as increased or decreased discussion levels on weekends, in order to isolate the movements that signal meaningful events.

(2) On Balance Volume

[0182] On Balance Volume (OBV) uses stock trading volume and price to quantify the level of buying and selling in a security. In a preferred embodiment of the present invention, OBV is used, e.g., by substituting the number of discussion participants for the stock volume. In this context, OBV is a negative indicator, i.e., when it is rising, price tends to fall; when it falls, price tends to rise.

(3) Moving Average Convergence-Divergence

[0183] Moving Average Convergence-Divergence (MACD) is a technical analysis that may be applied to the discussion levels in the communities. MACD generates signals by comparing short-term and long-term moving averages; the points at which they cross one another can be buy or sell signals, depending on their directions. MACD can signal when a community's discussion level rises above the recent averages, which is often an indicator of rising nervousness.

Link Analysis

[0184] In one embodiment of the present invention an "80/20 rule," supported by social network research, is used wherein only the 20 percent of participants whose posts are "closest" (in time) to significant objective data inflection points are analyzed. While this method simplifies the task of analyzing the data, there is some risk that opinion-leading groups may be overlooked. Such groups comprise

individuals that do not consistently post at the same time, but as a group exhibit the characteristics of individual opinion leaders. For example, it is possible Bob, Sam and George form a positive opinion leader group, i.e., when any one of them posts a message, prices tend to rise. Data mining link analysis tools are used to explore for these kinds of relationship and to identify groups of local pseudonyms whose behavior as a group exhibits predictive patterns.

Geographic Visualization

[0185] Tools for geographic visualization display the distribution of information on a map. Although geographic location is unknown for many of the local pseudonyms being monitoring, it is available for some of them and will be tracked as the information becomes available. This analysis allows monitoring of the awareness of a topic, such as a newly released consumer media device, as it spreads throughout the United States and other countries. This analysis will help marketers decide where promotional and advertising budgets can be spent most effectively. Marketing experience and the mathematics of social networks predict that awareness follows a stair-step pattern. The analysis results of the present invention can be used to identify these plateaus very early, allowing marketers to cut spending earlier than they otherwise would.

(4) Clustering

[0186] Cluster analysis allows discovery of groups of local pseudonyms that “travel in the same circles.” For example, there may be a group of 20 local pseudonyms that tend to participate in discussions on five topics. This cluster of shared interests is a means of automatically discovering that there is some kind of relationship among the five topics. In the financial market, it implies that people who are interested in any

one of the five companies are likely to find the other four interesting. Presenting these as recommendations is a form of *collaborative filtering*, because it helps the user select a few new topics of interest out of thousands of possibilities. The most significant aspect of this analysis is that the computer system needs no knowledge of *why* the topics are related; the system can therefore discover new relationships.

(5) Regression

[0187] Regression analysis is a well-known method of correlating sets of data. Regression is the most fundamental means for identifying if the patterns in communities have a positive, negative or insignificant correlation to external events.

(6) Neural Networks and Genetic Algorithms

[0188] Neural networks and genetic algorithms are machine-learning approaches for finding optimal solutions to complex problems. Neural nets take a set of inputs, which might be various parameters about a community, such as message level, ratio of positive to negative opinion leaders, etc., and discover relative weightings to achieve a desired outcome, such as a predicted stock price. Neural nets have been used successfully in other types of financial forecasting and analysis. Genetic algorithms evolve solutions to complex problems by imitating the competitive nature of biological genetics. Factors under consideration must be encoded in a binary form and a system for ranking the value of the outcome is created. Software applications used to perform such analyses in the present invention are commercially available from, e.g., Ward Systems Group, Inc. of Frederick, MD.

Report Generation Subsystem

[0189] Report presentation subsystem 36 extracts the results of the analysis performed by analysis subsystem 34 for presentation to end-users 9. In a preferred

embodiment, report generation subsystem 36 and presents it to end-users via a Web-based user interface. In this embodiment, the reports are published using a variety of formats, such as, e.g., PDF, HTML, and commercially available spreadsheets or word processors, and the like. End-users 9 may use any suitable Web browser to view and receive the reports generated by report generation subsystem 36. Examples of such Web browsers are available from Netscape, Microsoft, and America Online. In an alternative embodiment, report generation subsystem 36 presents the results in written reports that may be printed and distributed.

[0190] Report generation subsystem 36 produces and displays some reports automatically and other reports may be specifically requested by end-users 9. For example, in a preferred embodiment, dynamic content boxes are automatically generated and displayed via a Web server. Such dynamic content boxes may include a report on the current market mood, displaying a visual indicator for the NASDAQ 100, for example. Such a market mood graph may contain the NASDAQ 100 market mood over the last 1 year together with the closing price of the NASDAQ 100 for the same period. Another dynamic content box could, e.g., display the top five companies where activity is spiking the greatest over the last 1 day versus activity recorded over the last 10 days. Alternatively, the dynamic content box could display the top five companies that are being discussed by the top five buy signalers. Other such reports can be generated and displayed automatically such that when end-users 9 connect to the Web server, the reports are presented without the need for requesting the information.

[0191] Other reports that may generated by report generation subsystem 36 include for example, a list of the most recent subjects posted by the top buy signaler for each of the top five most positive market mood companies and real-time trends such as information about postings to Internet based communities. These reports and other may be dynamically built by report generation subsystem 36 based on requests for information from end-users 9. For example, end-user 9 may specify a community, a local or universal pseudonym or a topic about which detailed information can be presented. For example, if an end-user requests a report concerning pseudonyms (local or universal) meeting a certain criteria, report generation subsystem 36 executes a search of all matching pseudonyms together with the source of the pseudonym (Yahoo, Raging Bull, etc.), if local, and links to a profile page for each pseudonym.

[0192] A local pseudonym's profile page comprises another report generated by subsystem 36 and includes, e.g., the local pseudonym and its source; an e-mail address of the local pseudonym on the community, if one exists; the total number of posts that the local pseudonym has made in discussion groups that are being tracked; the number of different topics that the local pseudonym has posted to in discussion groups that are being tracked; the most recent posting date that the local pseudonym has made to any discussion group and a link to that posting; a list of most recent postings to discussion groups categorized by topics; the local pseudonym's reputation score for each category; a graphical representation of the local pseudonym's reputation (e.g., Figure 4); and the like.

[0193] In addition to retrieving reports concerning particular local pseudonyms, report generation subsystem 36 allows end-users 9 to locate detailed information about each topic (company, book, movie, etc.). For example, if an end-user requests a report on a particular company, by e.g., the stock symbol or the company name, another search is executed. Report generation subsystem 36 displays information such as a list of all matching companies; the name of the company; the stock symbol of the company; and a link to a company profile page where users can obtain detailed information about that particular company.

[0194] A company profile is similar to a pseudonym's profile page. That is, the company profile page is another report generated and displayed by report generation subsystem 36. In a preferred embodiment, the company profile page comprises detailed information about a particular company, especially information that relates to postings in stock message forums for that company. Other information that may be displayed includes, e.g., the name of the company; the stock exchange that the company is a member of; the domain name for the company's home page and a link; a link to the company's stock board on Yahoo, Raging Bull, Motley Fool or other prominent electronic discussion forums; a list of the most frequent posters on the company's stock discussion groups; the top buzz accelerators and the top buzz decelerators for the company's stock discussion groups; and top buy and sell signalers for the company's stock discussion groups.

[0195] For other topics, analogous profile pages can be presented. For example, a movie's profile page may comprise the movie's name, the producer, and other

objective information as well as identification of the top buzz accelerators and decelerators, and other results of output by analysis subsystem 34.

UNIVERSAL PSEUDONYM REGISTRATION SYSTEM

[0196] As shown in Figure 5, the present invention may include universal pseudonym registration system 40. Universal pseudonym registration system 40 allows end-users, such as end-users 41 to sign-up (or register) for universal pseudonym services. The services include creation of universal pseudonyms for use in posting messages to electronic discussion forums; the capability to build a reputation in a community through persistent universal pseudonym identity, opt-in marketing services (wherein universal pseudonyms can be registered to receive selected categories of marketing information). For example, an end-user can register one universal pseudonym and specify an interest in comic books, and register another universal pseudonym with an interest in stock market forecasts. Although the two universal pseudonyms belong to the same person, the person can more easily differentiate and select the type of information sought at a particular moment. Moreover, registration with universal pseudonym registration system 40 provides a means for end-users 41 to provide certain demographic information (age, gender, salary, and the like) without revealing their actual identity.

[0197] In a preferred embodiment, universal pseudonym registration system 40 provides a digital signature that registered universal pseudonyms may use to prove their identity as a registered universal pseudonym. The digital signature allows the user to indicate within a message posting that the local pseudonym is linked to other pseudonyms via a universal pseudonym that can be verified by universal pseudonym

registration system 40. In this manner, not only can the system and method of the present invention track the user's posting on various communities to rate the user's reputation across multiple communities, it also informs other community participants that the user has registered the local pseudonym on universal pseudonym registration system 40. As discussed above, a user may be known by the local pseudonym john@yahoo.com in the Yahoo.com community, and by the local pseudonym john2@amazon.com in the Amazon.com community. In this case, the end-user can register both local pseudonyms with universal pseudonym registration system 40 and associate the two local pseudonyms with a single universal pseudonym, e.g., john.doe@pseud.org. When positing messages under either local pseudonym, the end-user authenticates his or her identity by providing the digital signature in the message. When other participants in the community see the digital signature, they can verify that the end-user john@yahoo.com is the same end-user john2@amazon.com by checking universal pseudonym registration system 40.

[0198] Universal pseudonym registration system 40 is a useful addition to the overall operation of the system and method of the present invention. By allowing end-users to select a universal pseudonym and associate various local pseudonyms, the data collected and analyzed can have more points for correlation. End-users are benefited both by better analysis results and by more control over their personal identifying information.

[0199] The foregoing disclosure of the preferred embodiments of the present invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed.

Many variations and modifications of the embodiments described herein will be obvious to one of ordinary skill in the art in light of the above disclosure. The scope of the invention is to be defined only by the claims appended hereto, and by their equivalents.

[0200] Further, in describing representative embodiments of the present invention, the specification may have presented the method and/or process of the present invention as a particular sequence of steps. However, to the extent that the method or process does not rely on the particular order of steps set forth herein, the method or process should not be limited to the particular sequence of steps described. As one of ordinary skill in the art would appreciate, other sequences of steps may be possible. Therefore, the particular order of the steps set forth in the specification should not be construed as limitations on the claims. In addition, the claims directed to the method and/or process of the present invention should not be limited to the performance of their steps in the order written, and one skilled in the art can readily appreciate that the sequences may be varied and still remain within the spirit and scope of the present invention.